WE CLAIM:

1. A method for identifying the misuse of authorized access to a digital data gathering system by a user, comprising:

a) constructing a user cluster index for a user of a digital data gathering system;

wherein the user cluster index comprises a list of families of data to which data from digital data gathering results of the user were categorized;

b) monitoring families of the digital data gathering results of the user;

c) comparing the families of the digital data gathering results of the user to the user cluster index to determine anomalies in the digital data gathering results; and

d) identifying a potential misuse when an anomaly is detected.

2. The method for identifying the misuse of authorized access to a digital data gathering system by a user according to Claim 1, further comprising:

a) comparing the anomalies to the user cluster index to determine the ratio of anomalies to existing clusters; and

b) reporting a potential misuse when the ratio exceeds a predetermined threshold.

3. The method for identifying the misuse of authorized access to a digital data gathering system according to Claim 1, further comprising:

a) monitoring digital data gathering results of the user;

b) constructing a user lexicon for a user of a digital data gathering system;

wherein the user lexicon comprises a list of words or phrases gathered from documents of the digital data gathering results of the user;

c) comparing words or phrases gathered from the documents of the digital data gathering results to the user lexicon to determine anomalies in the digital data gathering results; and

d) identifying a potential misuse when an anomaly is detected.

4. The method for identifying the misuse of authorized access to a digital data gathering system according to Claim 3, further comprising:

a) monitoring digital data gathering queries of the user;

b) and wherein the user lexicon further comprises a list of words or phrases gathered from the monitoring of the queries;

c) comparing queries of the user to the user lexicon to determine anomalies in the queries; and

d) identifying a potential misuse when an anomaly is detected.

5. The method for identifying the misuse of authorized access to a digital data gathering system by a user according to Claim 3, further comprising:

a) determining a ratio of anomalies to words or phrases in the lexicon; and

b) reporting a potential misuse when the ratio exceeds a predetermined threshold.

6. The method for identifying the misuse of authorized access to a digital data gathering system by a user according to Claim 3, wherein the user lexicon comprises a list of words or word strings identifying particular words or types of words, or both, extracted from documents returned in response to user queries.

7. The method for identifying the misuse of authorized access to a digital data gathering system according to Claim 1, further comprising:

a) constructing a structured data profile for a user of a digital data gathering system;

b) wherein the structured data profile comprises a list of data identifying workplace characteristics of the user;

c) comparing the digital data gathering results of the user to the structured data profile to determine whether the digital data gathering results are congruent with the structured data profile; and

d) identifying a potential misuse when the digital data gathering results are not congruent with the structured data profile.

8. The method for identifying the misuse of authorized access to a digital data gathering system according to Claim 7, further comprising:

a) the structured data profile comprising a structured data profile lexicon of terms and phrases indicating valid user activity; and

b) identifying a potential misuse when the digital data gathering results are not congruent with the structured data profile.

9. The method for identifying the misuse of authorized access to a digital data gathering system according to Claim 3, further comprising:

a) constructing a structured data profile for a user of a digital data gathering system;

b) wherein the structured data profile comprises a list of data identifying workplace characteristics of the user;

c) comparing the digital data gathering results of the user to the structured data profile to determine whether the digital data gathering results are congruent with the structured data profile; and

d) identifying a potential misuse when the digital data gathering results are not congruent with the structured data profile.

10. A method for identifying the misuse of authorized access to a digital data gathering system by a user, comprising:

a) monitoring digital data gathering results of the user;

b) constructing a user lexicon for a user of a digital data gathering system;

wherein the user lexicon comprises a list of words or phrases gathered from documents of the digital data gathering results of the user;

c) comparing words or phrases gathered from the documents of the digital data gathering results to the user lexicon to determine anomalies in the digital data gathering results; and

d) identifying a potential misuse when an anomaly is detected.

11. The method for identifying the misuse of authorized access to a digital data gathering system by a user according to Claim 10, further comprising:

a) determining a ratio of anomalies to words or phrases in the lexicon; and

b) reporting a potential misuse when the ratio exceeds a predetermined threshold.

12. The method for identifying the misuse of authorized access to a digital data gathering system by a user according to Claim 10, wherein the user lexicon comprises a list of words or word strings identifying nouns extracted from documents returned in response to user queries.

13. The method for identifying the misuse of authorized access to a digital data gathering system according to Claim 10, further comprising:

a) constructing a structured data profile for a user of a digital data gathering system;

b) wherein the structured data profile comprises a list of data identifying workplace characteristics of the user;

c) comparing the digital data gathering results of the user to the structured data profile to determine whether the digital data gathering results are congruent with the structured data profile; and

d) identifying a potential misuse when the digital data gathering results are not congruent with the structured data profile.

14. A method for identifying the misuse of authorized access to a digital data gathering system by a user, comprising:

a) constructing a structured data profile for a user of a digital data gathering system;

b) wherein the structured data profile comprises a list of data identifying workplace characteristics of the user;

c) monitoring digital data gathering results of the user;

d) comparing digital data gathering results of the user to the structured data profile to determine whether the digital data gathering results are congruent with the structured data profile; and

e) identifying a potential misuse when the digital data gathering results are not congruent with the structured data profile.

15. A method for identifying the misuse of authorized access to an information retrieval system by a user, comprising:

a) constructing a profile of use for a user of an information retrieval system;

b) the profile including a user lexicon of user result terms, a user cluster index of result document categories, and a structured data profile of known user characteristics;

c) monitoring information retrieval results of the user;

d) comparing the information retrieval results of the user to the user profile to determine the anomalies in the new queries and results;

e) identifying a potential misuse when an anomaly is detected;

f) comparing the information retrieval results of the user to the structured data profile to determine whether the new query terms and results are congruent with the structured data profile; and

g) identifying a potential misuse when the information retrieval results are not congruent with the structured data profile.

16. The method for identifying the misuse of authorized access to an information retrieval system by a user according to Claim 15, further comprising:

a) determining a ratio of anomalies to words or phrases in the lexicon; and

b) reporting a potential misuse when the ratio exceeds a predetermined threshold.

17. The method for identifying the misuse of authorized access to an information retrieval system according to Claim 15, further comprising: weighting potential misuses identified from the user lexicon, the user cluster index, and the structured data profile to determine a report of potential misuse.

18. The method for identifying the misuse of authorized access to an information retrieval system according to Claim 15, further comprising: sending a notification of potential misuse when a potential misuse is identified from two or more of the user lexicon, the user cluster index, and the structured data profile.

19. The method for identifying the misuse of authorized access to an information retrieval system according to Claim 15, wherein the user lexicon comprises a list of words or phrases gathered from metadata of documents returned in the query results.

20. The method for identifying the misuse of authorized access to an information retrieval system according to Claim 15, wherein the user lexicon comprises a list of words, or types of words, or both, extracted from documents returned in the query results.

21. The method for identifying the misuse of authorized access to an information retrieval system according to Claim 15, wherein the user cluster index comprises a list of families of data to which the data of the user information retrieval results have been categorized.

22. A method for detecting misuse by a user of an information retrieval system having a document collection, comprising the steps of:

a) pre-clustering the document collection;

b) tracking the cluster from which any document read by the user originates;

c) building up a profile of the user based on most frequently accessed clusters over a time sufficient to establish a confidence threshold for validity of the profile of the user;

d) tracking each time the user retrieves and reads a document outside of the most frequently accessed clusters; and

e) establishing a misuse threshold number for documents read outside of the most frequently accessed clusters and after the misuse threshold number is obtained, signaling that a potential misuse may have occurred.

23. A method for detecting misuse by a user of an information retrieval system having a document collection, comprising the steps of:

a) retrieving documents in response to user queries;

b) clustering the retrieved documents by category;

c) establishing and obtaining a threshold number of retrieved documents and after the threshold number of retrieved documents is obtained, determining a size for each clusters, and further denoting clusters of a large enough size as valid clusters; and

d) determining if a sufficient number of retrieved documents do not participate in any valid cluster and if not, sounding an alarm.

24. A method for detecting misuse by a user of an information retrieval system having a document collection, comprising the steps of:

a) identifying top weighted terms from documents retrieved by the user and storing the top weighted terms in a user-specific lexicon;

b) tracking user activity until the rate of new terms added slows and the user-specific lexicon stabilizes to form a user profile;

c) identifying for each new query, if the top weighted terms are in the user-specific lexicon;

d) tracking a ratio of newly occurring terms to existing user-specific lexicon terms; and

e) if the ratio of newly occurring terms to existing user-specific lexicon terms exceeds a threshold, sending an alarm.

25. The method for detecting misuse by a user of an information retrieval system having a document collection, according to Claim 24, further comprising the steps of:

a) tagging the documents to identify words in the documents by type;

b) running an original query of terms and phrases;

c) selecting specific types of words from relevant documents retrieved by the original query and adding these terms to a second query; and

d) iteratively selecting specific types of words from relevant documents retrieved by each query and adding the selected specific types of words to a further query to filter the user-specific lexicon.

26. A method for detecting misuse by a user of an information retrieval system having a document collection, comprising the steps of:

a) identifying structured data sources that can be used to identify what the user is working on;

b) querying these sources and, for each source, mapping a structured result into a structured data lexicon of terms and phrases that indicate valid user activity;

c) for each new query, tracking a ratio of terms found in the structured data lexicon to those not found in the structured data lexicon; and

d) if the ratio exceeds a threshold, sending an alarm that a misuse may have occurred.

27. A method for detecting misuse by a user of an information retrieval system having a document collection, comprising the steps of:

a) identifying structured data sources that can be used to identify what the user is working on;

b) querying the identified structured data sources and, for each source queried, mapping a structured result into a structured data lexicon of terms and phrases that indicate valid user activity;

c) for each new query, retrieving relevant documents for that new query;

d) extracting key terms from the relevant documents;

e) identifying the ratio of key retrieved terms found in the lexicon to those not found in the lexicon; and

f) if the ratio exceeds a threshold, sending an alarm that a misuse may have occurred.